Can Automated Gesture Recognition Support the Study of Child Language Development?

Soumitra Samanta, Colin Bannard, Julian Pine and The Language05 Team
{soumitra.samanta, colin.bannard, julian.pine}@liverpool.ac.uk
Department of Psychological Sciences
University of Liverpool, Liverpool, UK

Abstract

Children's prelinguistic gestures play a central role in their communicative development. Early gesture use has been shown to be predictive of both concurrent and later language ability, making the identification of gestures in video data at scale a potentially valuable tool for both theoretical and clinical purposes. We describe a new dataset consisting of videos of 72 infants interacting with their caregivers at 11&12 months, annotated for the appearance of 12 different gesture types. We propose a model based on deep convolutional neural networks to classify these. The model achieves 48.32% classification accuracy overall, but with significant variation between gesture types. Critically, we found strong (0.7 or above) rank order correlations between by-child gesture counts from human and machine coding for 7 of the 12 gestures (including the critical gestures of declarative pointing, hold outs and gives). Given the challenging nature of the data - recordings of many different dyads in different environments engaged in diverse activities - we consider these results a very encouraging first attempt at the task, and evidence that automatic or machine-assisted gesture identification could make a valuable contribution to the study of cognitive development.

Keywords: Deep learning, child gesture recognition, language development.

Introduction

The last two decades have seen a renewed research focus on the production of prelinguistic gestures as a critical component of communicative development. Tomasello, Carpenter, and Liszkowski (2007), for example, argue that in children's early pointing behavior we can already see the cognitive tools on which the subsequent development of conventional linguistic communication will hinge. Colonnesi, Stams, Koster, and Noom (2010) perform a meta-analysis of 25 different studies and find robust evidence of a link between early pointing and later language ability.

Other papers have focused on the importance of other kinds of gestures in early communicative development. For example, Boundy, Cameron-Faulkner and Theakston (2019) report evidence of the intentionally communicative nature of hold-out gestures (holding out objects in order to direct others' attention), while Cameron-Faulkner et al. (2015) report evidence that such gestures call on the same cognitive abilities as pointing. McKean et al. (2016) report evidence that this gesture type is a valuable predictor of later language ability.

Donnellan et al. (2019) report an analysis of a cohort of children in which they evaluate the predictive value with regard to later vocabulary of a very large set of gestural and vocal behaviors. They find that a number of gesture types have predictive value. However, they also find that the gesture types vary greatly in their value, and even in their direction of effect, with the rate of some gesture types being a positive predictor of later vocabulary (the more a child produces that gesture, the larger their subsequent vocabulary) but others being a negative predictor (the more a child produces that gesture, the smaller their subsequent vocabulary). This highlights the importance of distinguishing gesture types from one another. It is this objective to which the current work makes a contribution.

All of the work described above has involved the coding of gesture types in videos of child-caregiver interaction. This work has all been done manually. This is very time consuming and expensive to perform, a factor that restricts progress in the field, and limits the possibility of application to clinical contexts - assessing an individual child's gestural development might be of considerable clinical utility, but is not a practical proposition if it requires the manual analysis of hours of video.

In this paper, we propose an automatic child gesture recognition method to support the analysis of child gesture. Our method is based on a deep convolutional neural network (DCNN) model combined with Support Vector Machines. We formulate the child gesture recognition problem as a human action recognition problem. We focus on two different types of information: *shape* and *motion*. We trained a spatial DCNN to capture the *shape* of the actor and a motion DCNN to capture the *motion* information. We then pooled these two types of features from all the frames and combined them to give the final feature representation. We classify each video representation using a linear Support Vector Machine (SVM). We evaluate our model on a child language development dataset from Rowland, Durrant, Peter, Bidgood, and Pine (2015b) and report promising results. In this paper, our contributions are as follows:

- We propose an automatic child gesture recognition method from video, which has wide application in cognitive science communities.
- We explore the application of the deep learning (two-stream DCNN) to a new type of action recognition.
- We introduce a new human gesture/action recognition

dataset that will be of considerable interest to the cognitive science communities.

Related work

To the best of our knowledge, there has been no previous work on automatic child gesture recognition (CGR). Gesture recognition in the context of communication has mostly involved adult sign language recognition. Starner, Weaver and Pentland (1998) used hidden Markov models to create an American Sign Language (ASL) recognition system from single fixed camera hand gesture videos and head mounted camera videos. Farhadi and Forsyth (2006) proposed a discriminative word model for ASL alignment from the video transcript. Later Farhadi, Forsyth and White (2007) used transfer learning to extend a model of ASL from artificial data (an avatar signer) to human data. They calculated different features (position, orientation and velocity of the hands and head and their SIFT features representation) for each video frame and concatenated *seven* consecutive frames as a final feature representation. Their model relies heavily on the handcrafted features. Nayak et al. (2009) represent ASL as multidimensional time series data and extract the stable part of the sign (called signemes) from multiple sentences using Iterated Conditional Modes.

Some early work (Buehler, Everingham, & Zisserman, 2009) on the automatic learning of British Sign Language (BSL) from TV broadcasting explored the pictorial structure model (Felzenszwalb & Huttenlocher, 2005). Buehler, Everingham and Zisserman (2009) first estimated the human upper body part (shoulder, arms and hand) configurations using a stochastic search method. They then used multiple instance learning to align the English words with the corresponding BSL signs. Later, Pfister et al. (2014) treated the pose estimation problem as a regression problem using a deep convolutional neural network (DCNN) for BSL recognition. Their DCNN consists of five convolutional layers followed by three fully connected layers. After each convolutional layer they normalized the convolution response and then pooled. Pfister, Charles and Zisserman (2014) proposed a domain adaptation based method for BSL and Italian hand gestures recognition. They used a Global Alignment Kernel (Cuturi, 2011) to overcome the alignment problem in Dynamic Time Warping (DTW) (Sakoe & Chiba, 1978). There is also work on other signed languages (Holden, Lee, & Owens, 2007; Zhang, Zhou, Xie, Pu, & Li, 2016; Hore et al., 2017)

Our model is formally similar to the human action recognition model of Simonyan and Zisserman (2014). They trained two different DCNNs: one for the spatial features and another for the temporal features, and used different types of fusion techniques for both the features to classify different human action. This two-stream-based DCNN and its different variants (Zha, Luisier, Andrews, Srivastava, & Salakhutdinov, 2015; Wang et al., 2016; Yu, Wang, Huang, Yang, & Xu, 2016; Crasto, Weinzaepfel, Alahari, & Schmid, 2019) are the most successful methods to-date for human action recognition.

Proposed Method

Our child gesture recognition method consists of two steps: i) video feature representation and ii) labeling of each video frame using Support Vector Machines (SVMs). The proposed method is different from Simonyan and Zisserman (2014) in two ways: i) we use a temporal pooling on spatial and temporal feature for our final video feature representation and ii) instead of combining the SVM scores from two network, we train the SVM on combined features from two networks. We will first describe our feature representation.

Video feature representation

For video feature representation, we use a deep convolution neural network (DCNN) based on the spatial and temporal network in Simonyan and Zisserman (2014). We trained two different networks: 1) a *spatial net*, to capture the shape information of the actor and 2) a *temporal net*, to capture the motion information of the child gestures. From both the DCNNs, we take the last layer output (before the classification layer) as our spatial and motion feature representation. We concatenate these feature vectors to get the final video representation. Figure 1 shows our feature representation.

For the spatial net, we use a DCNN similar to the ResNet 101 (He, Zhang, Ren, & Sun, 2016) architecture with a 3-channel (RGB) video frame as input and $C(=12)$ different gesture classes as output. We randomly select 10 frames from a fixed size video clip of 30 frames. We accumulate the classification score (using $Softmax$) from all the 10 frames and calculate the classification error using a *cross-entropy* loss function.

For each video clip, we calculate the displacement of each pixel (the optical flow; Bruhn, Weickert, & Schnorr, 2005) between two consecutive frames. We then randomly select 10 consecutive frames from the fixed size video clip of size 30 frames. We stack the optical flow (both horizontal and vertical one by one) and make a three-dimensional matrix of size $20 \times r \times c$, where r and c are the height and width of the video frame, respectively. We use a similar DCNN to the *spatial net* (ResNet 101; He et al., 2016) architecture with 20-channel as input for our *Temporal net*.

Take a video clip V (a particular child gesture) divided into m sub-clips $V_1, V_2, ...V_m$. For our spatial and temporal networks, we use V_i as input and get a high-level d_{sn} and d_{tn} dimensional spatial and temporal feature representation $f_{sn,i}$ and $f_{tn,i}$, $i = 1,2,...,m$, respectively.

As different child gestures have different durations and we need to represent each gesture with a fixed dimension vector, we use a temporal pooling over all the feature representation on both spatial $f_{sn_1}, f_{sn_2}, ..., f_{sn_m}$ and temporal $f_{tn_1}, f_{tn_2}, ..., f_{tn_m}$ feature representation from that clip. Let h be the pooling operator, then the vector z_{sn} and z_{tn} represent the spatial and temporal feature representation of V defined as:

$$z_j = h(\{f_{j_1}, f_{j_2}, ..., f_{j_m}\}); j \in \{sn, tn\} \quad (1)$$

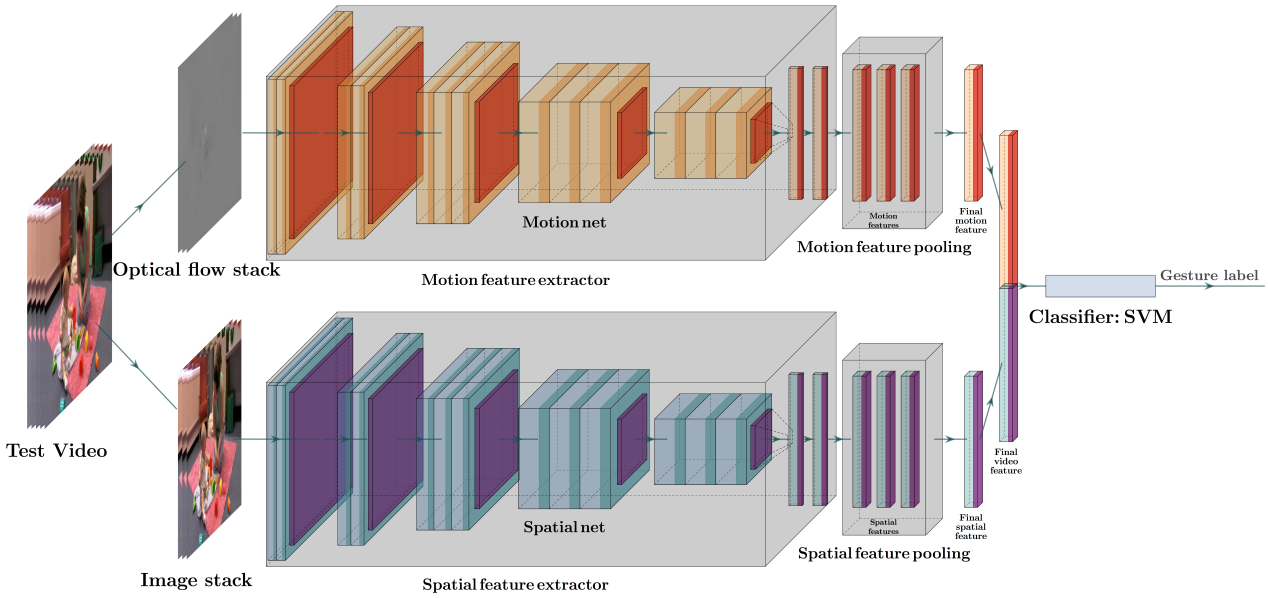


Figure 1: Proposed model architecture

We concatenate z_{sn} and z_{tn} to get the final representation $z = [z_{sn}, z_{tn}]$ of the video V.

There are different types of feature pooling operator in the literature (Boureau, Ponce, & LeCun, 2010). Max and average pooling (Murray & Perronnin, 2014; Yang, Yu, Gong, & Huang, 2009) are the most popular choices for visual recognition. We use h as a max-pooling operator for our video feature representation, which is defined as:

$$z_{sn,i} = max(\{f_{sn_1,i}, f_{sn_2,i}, ..., f_{sn_m,i}\}); i = 1, 2, ..., d_{sn} \quad (2)$$

$$z_{tn,i} = max(\{f_{tn_1,i}, f_{tn_2,i}, ..., f_{tn_m,i}\}); i = 1, 2, ..., d_{tn} \quad (3)$$

We classify each child gesture based on its feature representation vector $z \in \mathbb{R}^{d_{sn}+d_{tn}}$ by concatenation of z_{sn} and z_{tn} (see equations 2 and 3).

Labeling of each video frame using classification

During training, for each video clip we have a feature vector $z \in \mathbb{R}^{d_{sn}+d_{tn}}$. Let there be N training gesture clips. So we have a set of feature vectors $\{z^1, z^2, ..., z^N\}$, where each z^i represents a particular gesture class. We train a linear Support Vector Machine (SVM) with fixed cost parameter value ($c = 100$) using a one-vs-rest strategy.

Here, our main goal is to classify each frame of a long video, which contains multiple gestures. To label a particular video frame, we consider the consecutive $\tau - 1$ frames before that frame (total τ frames) and calculate the feature representation as described in the previous subsection. We then classify that feature using our trained SVM and consider that class label as the label of that frame. In this way, we label all the frames of that video. We evaluate our method using the dataset from (2015b) described in the next section.

Experimental Evaluation

Dataset

For our experimental evaluation, we have used the dataset from Rowland et al. (2015b). This dataset, which will shortly be publicly available for research purposes (2015a), contains videos of 72 children at two ages (11 & 12 months) engaging in various activities with their caretakers. The dataset also contains data at each age from the UK-CDI - an adaptation of the MacArthur Bates Communicative Inventory (Fenson et al., 2007), a widely used parental questionnaire that tells us about the child's gesture production along with other aspects of their communicative development. Most notably for our purposes it contains caregiver ratings of their child's use of a series of common gestures on a three-point scale (not yet, sometimes or often). The dataset was collected in home environments using a handheld moving camera. The camera parameters (pan, tilt, and zoom) were adjusted according to the infant movement in the room (both natural and artificial light sources). For each age group, there is an average of 30 mins of videos. For all the videos, child gestures have been manually coded by trained research assistants and cross-checked by the different students. They identified 12 different child gestures: *grasp object (GO), give (GV), hold out (HO), lower object (LO), object manipulation (OM), other (OT), point-declarative (PD), point-imperative (PI), reaches-imperative (RI), reaches-declarative (RD), retract object (RO),* and *share orientation (SO)*. Table 1 shows the number of occurences of each gesture class, respectively. In total, the dataset contains 22082 videos. For our evaluation, we have used a three-fold cross validation strategy. Sample images for each gesture are shown in Figure 2.

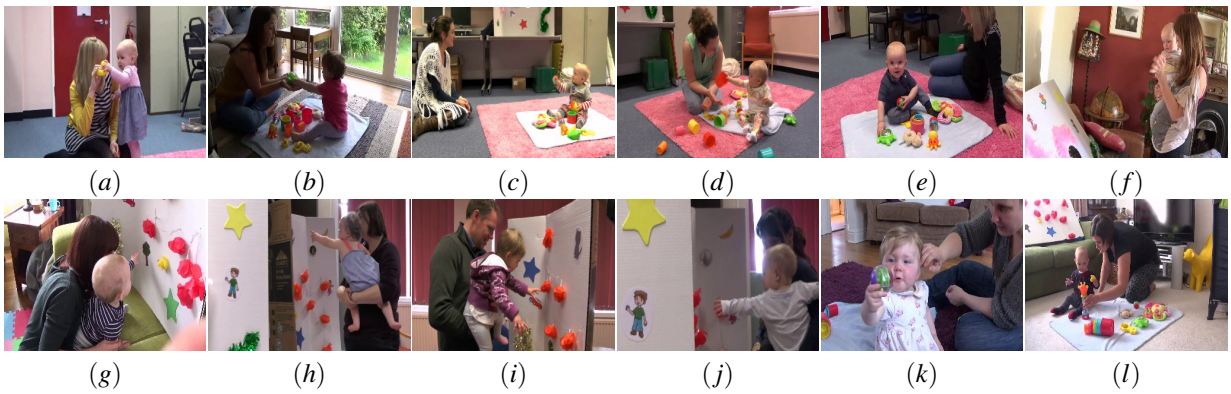


Figure 2: Sample frames of different child gestures in dataset: (a) grasp object (GO), (b) give (GV), (c) hold out (HO), (d) lower object (LO), (e) object manipulation (OM), (f) other (OT), (g) point-declarative (PD), (h) point-imperative (PI), (i) reaches-imperative (RI), (j) reaches-declarative (RD), (k) retract object (RO), and (l) share orientation (SO).

Table 1: Number of videos per gesture class.

Age	Different gestures												Total
	GO	GV	HO	LO	OM	OT	PD	PI	RH-1	RH-2	RO	SO	
11M	627	221	396	69	5988	153	401	208	127	938	42	589	9759
12M	720	316	409	107	6044	391	876	229	338	1188	46	1659	12323
All	1347	537	805	176	12032	544	1277	437	465	2126	88	2248	22082

Experimental setting

For each data partition, we trained our DCNN for video feature representation. We implemented our DCNN using the *PyTorch* library (Paszke et al., 2019). We used 500 epochs to train our networks with batch size 64. To update our network parameters, we use Adam update rules (Kingma & Ba, 2014) with learning rate 10^{-2} and decay after 100 epoch with a factor 10^{-1}. For optical flow calculation, we used *Coarse2Fine* methods (Brox, Andreś, Papenberg, & Weickert, 2004; Bruhn et al., 2005; *pyflow*, 2017). To reduce the displacement error, we capped the optical flow vector at [-20, 20], which is suitable for our *Temporal net* DCNN. Similar to Simonyan and Zisserman (2014), we stack both the horizontal and the vertical optical flow one-by-one to form our input to the *Temporal net*.

For our evaluation metric, we first calculate the gesture recognition accuracy (%) with respect to the hand coding. To give a measure of the practical utility of our method (in ranking each child's rate of production for each gesture relative to their peers), we report the *rank order correlation* (Spearman) over all children between the rate of each gesture according to the hand coding and the machine coding. We report these evaluations separately at each age (11M & 12M). Our test code is publicly available at `https://github.com/soumitrasamanta/child_gesture`.

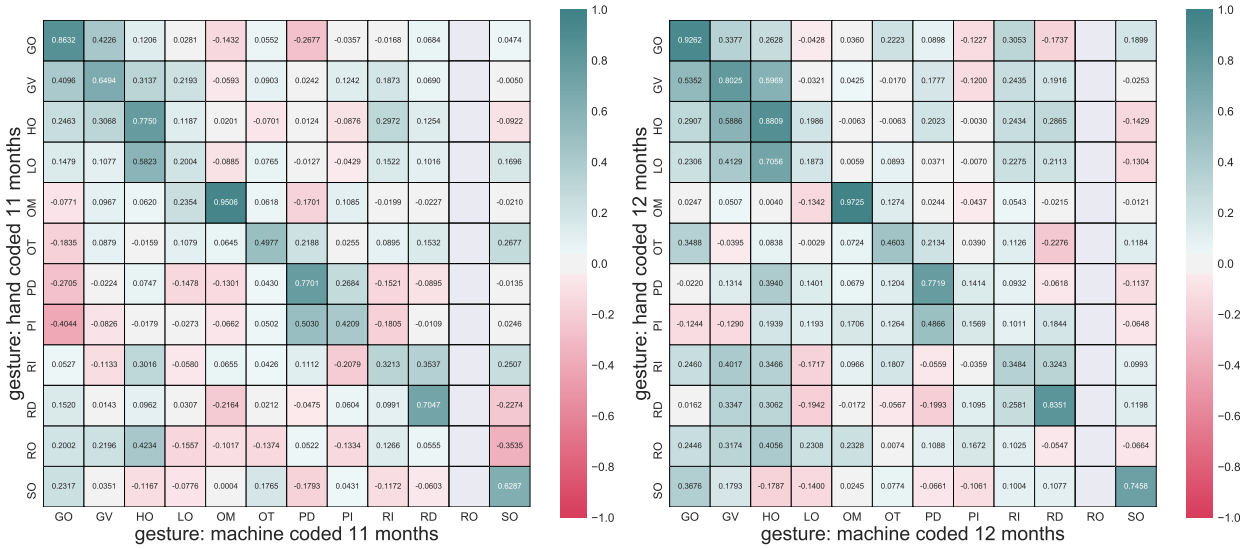
Experimental results

To see the individual feature (*spatial* and *temporal*) performance and their combination, we calculate the average accuracy over all the three-fold data partitions and the results are shown in the Table 2. From the Table 2, we see that the *motion* feature gives better performance than the *spatial* feature as in general human action recognition (Simonyan & Zisserman, 2014). The combination of both the features gives the highest performance of 48.32% accuracy.

Table 2: Average accuracy for different feature combinations

Feature type	Avg. accuracy (%)
spatial feature (SF)	43.92
motion feature (MF)	46.72
SF + MF	48.32

Figure 3 (a) and (b) shows the correlation between hand coded and machine coded gesture class for the 11M and 12M age groups, respectively. In Figure 3 (a) and (b), the horizontal and vertical axes shows the different machine coded and hand coded gestures, respectively. Each cell (i, j) ($i, j \in \{12$ different gestures$\}$), indicates the correlation between the i^{th} machine coded and j^{th} hand coded gesture. The diagonals show there is a strong correlation (0.7 or above) between machine coding and hand coding for each 7 individual gesture (GO, GV, HO, OM, PD, RD and SO). And critically there are considerably lower correlations across different gesture types (e.g. between machine coding for one gesture type and hand coding for another type), indicating that we are picking up gesture-specific information and not simply an overall rate of gesturing. Please note that due to the small number of



(a) (b)

Figure 3: Rank order correlation (Spearman) matrix between hand coded and machine coded labels for each age group (11M & 12M)

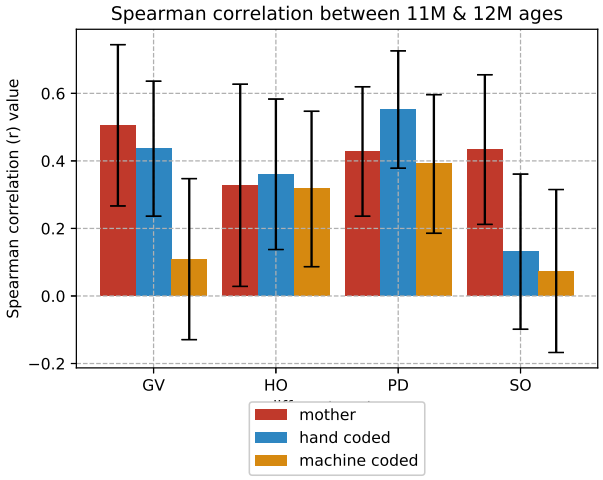


Figure 4: Individual rank order correlation between two age groups (11 & 12 month) of mother, hand coded and machine coded labels

data points for reach out gestures (Table 1), 0% accuracy was achieved and thus the RO column in Figure 3 is empty.

These results, then, suggest that the machine coding is giving us meaningful information about the rate of production for each gesture type for each child. In order to check this further we check the repeatability/test-retest reliability of the machine coding but checking the rank order correlation between the 11 and the 12 month data. For four gestures (GV, HO, PD, and SO), we have the mother's assessments (on a three-point scale) of the child's rate of producing that gesture type at each age taken from the UK-CDI. We report the correlation across the ages for these four gestures shown in Figure 4. For the critical behaviors of *hold out* and *point-declarative*, the machine coded correlation is similar to that observed for the mother and hand coded data.

Conclusion

We have proposed an automatic child gesture recognition method to aid the study and assessment of children's early communicative abilities - one with potential valuable clinical application. We have applied current state-of-the-art deep-learning-for-action-detection methods to this child gesture recognition problem. The primary real-world use of this method that we envisage is in assessing the gestures produced by a given child relative to other children, and we therefore evaluate our method by comparing the way it ranks the children for each gesture type to the ranking provided by hand coding. We found promising performance, particularly for the most important gesture types (declarative points, hold-outs and gives having been most consistently found to be associated with language development). This suggests that the automatic classification of gestures could be a valuable part of

the workflow in the analysis of communicative development. Our next step will be to combine this classification method with the first step of the analysis process - the detection of gestures in free video. If successful, this will provide a valuable end-to-end tool for use by researchers and practitioners.

References

Boundy, L., Cameron-Faulkner, T., & Theakston, A. (2019). Intention or attention before pointing: Do infants' early holdout gestures reflect evidence of a declarative motive? *Infancy*, *24*(2), 228-248.

Boureau, Y.-L., Ponce, J., & LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the ICML* (p. 111-118).

Brox, T., Andreś, Papenberg, B. N., & Weickert, J. (2004). High accuracy optical flow estimation based on a theory for warping. In *Proceedings of the ECCV* (p. 25-36).

Bruhn, A., Weickert, J., & Schnorr, C. (2005). Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *IJCV*, *61*(3), 211-231.

Buehler, P., Everingham, M., & Zisserman, A. (2009). Learning sign language by watching TV (using weakly aligned subtitles). In *Proceedings of the CVPR* (p. 2961-2968).

Cameron-Faulkner, T., Theakston, A., Lieven, E., & Tomasello, M. (2015). The relationship between infant holdout and gives, and pointing. *Infancy*, *20*(5), 576-586.

Colonnesi, C., Stams, G. J. J., Koster, I., & Noom, M. J. (2010). The relation between pointing and language development: A meta-analysis. *Developmental Review*, *30*(4), 352-366.

Crasto, N., Weinzaepfel, P., Alahari, K., & Schmid, C. (2019). MARS: Motion-augmented rgb stream for action recognition. In *Proceedings of the CVPR* (p. 7882-7891).

Cuturi, M. (2011). Fast global alignment kernels. In *Proceedings of the ICML* (p. 929-936).

Donnellan, E., Bannard, C., McGillion, M. L., Slocombe, K. E., & Matthews, D. (2019). Infants? intentionally communicative vocalizations elicit responses from caregivers and are the best predictors of the transition to language: A longitudinal investigation of infants? vocalizations, gestures and word production. *Developmental Science*, *0*(0), e12843. doi: 10.1111/desc.12843

Farhadi, A., & Forsyth, D. (2006). Aligning asl for statistical translation using a discriminative word model. In *Proceedings of the CVPR* (p. 1471-1476).

Farhadi, A., Forsyth, D., & White, R. (2007). Transfer learning in sign language. In *Proceedings of the CVPR* (p. 1-8).

Felzenszwalb, P. F., & Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *IJCV*, *61*(1), 55-79.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *Macarthur-bates communicative development inventories*. Brookes Publishing Co.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the CVPR* (p. 770-778).

Holden, E.-J., Lee, G., & Owens, R. (2007). Automatic recognition of colloquial australian sign language. In *Proceedings of the WACV* (p. 183-188).

Hore, S., Chatterjee, S., Santhi, V., Dey, N., Ashour, A. S., Balas, V. E., & Shi, F. (2017). Indian sign language recognition using optimized neural networks. In V. E. Balas, L. C. Jain, & X. Zhao (Eds.), *Information technology and intelligent transportation systems* (p. 553-563). Springer.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, *abs/1412.6980*.

McKean, C., Law, J., Mensah, F., Cini, E., Eadie, P., Frazer, K., & Reilly, S. (2016). Predicting meaningful differences in school-entry language skills from child and family factors measured at 12 months of age. *International Journal of Early Childhood*, *48*(3), 329-351.

Murray, N., & Perronnin, F. (2014). Generalized max pooling. In *Proceedings of the CVPR* (p. 2473-2480).

Nayak, S., Sarkar, S., & Loeding, B. (2009). Automated extraction of signs from continuous sign language sentences using iterated conditional modes. In *Proceedings of the CVPR* (p. 2583-2590).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the NeurIPS* (p. 8024-8035).

Pfister, T., Charles, J., & Zisserman, A. (2014). Domain-adaptive discriminative one-shot learning of gestures. In *Proceedings of the ECCV* (p. 814-830).

Pfister, T., Simonyan, K., Charles, J., & Zisserman, A. (2014). Deep convolutional neural networks for efficient pose estimation in gesture videos. In *Proceedings of the ACCV* (p. 538-552).

pyflow. (2017). `https://github.com/pathak22/pyflow`.

Rowland, C., Durrant, S., Peter, M., Bidgood, A., & Pine, J. (2015a). *Language05 dataset.* `https://archive.mpi.nl/`.

Rowland, C., Durrant, S., Peter, M., Bidgood, A., & Pine, J. (2015b). *The language 0-5 project.* `https://osf.io/kau5f/`.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. on ASSP*, *26*(1), 43-49.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Proceedings of the NIPS* (p. 568-576).

Starner, T., Weaver, J., & Pentland, A. (1998). Real-time american sign language recognition using desk and wearable computer based video. *IEEE Trans. on PAMI*, *20*(12), 1371-1375.

Tomasello, M., Carpenter, M., & Liszkowski, U. (2007). A new look at infant pointing. *Child Development*, *78*(3), 705-722.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Gool, L. V. (2016). Temporal segment networks: Towards

good practices for deep action recognition. In *Proceedings of the ECCV* (p. 20-36).

Yang, J., Yu, K., Gong, Y., & Huang, T. (2009, June). Linear spatial pyramid matching using sparse coding for image classification. In *Proceedings of the CVPR* (p. 1794-1801).

Yu, H., Wang, J., Huang, Z., Yang, Y., & Xu, W. (2016). Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the CVPR* (p. 4584-4593).

Zha, S., Luisier, F., Andrews, W., Srivastava, N., & Salakhutdinov, R. (2015). Exploiting image-trained cnn architectures for unconstrained video classification. In *Proceedings of the BMVC* (p. 1-13).

Zhang, J., Zhou, W., Xie, C., Pu, J., & Li, H. (2016). Chinese sign language recognition with adaptive HMM. In *Proceedings of the ICME* (p. 1-6).